The Truth in the Falsification of Artificial Intelligence

Abstract

The influence Karl Popper's falsificationist model has had on the scientific method and the demarcation problem is troublesome for the field of artificial intelligence (AI). According to Popper, the falsifiability of a hypothesis is a necessary condition for its scientific validity. Because the falsificationist model has been formative in the development of modern philosophy of science, it has become the primary way in which we demarcate the scientific from the non-scientific. However, as a consequence of our current, limited understanding of mental properties—such as intelligence, thought, and personal identity—I argue that it is unclear whether hypotheses concerning the design of artificial intelligence—particularly strong AI—are truly falsifiable. If this is the case, society's approach to and attitude towards future AI research and development regarding scientific methodology is in need of reevaluation. I conclude that we should either (1) aim to better define the concepts in philosophy of mind that we attribute to artificial intelligence to understand how they can be falsified in order to make quicker and more meaningful progress in AI or (2) recognize that there are aspects of AI that simply cannot be falsified and adapt our current scientific methodology to something more appropriate and inclusive than Popper's falsificationist principles.

Key words: artificial intelligence, falsificationism, Karl Popper, philosophy of science, philosophy of mind, cognitive science

The Truth in the Falsification of Artificial Intelligence

Research and development in the area of artificial intelligence is particularly interesting to technology-loving modern society. To computer scientists, cognitive scientists, philosophers of science and the mind, and laypersons alike, it is currently up for debate (1) exactly what artificial intelligence entails and (2) whether it is truly possible to design. For the following argument in this paper, artificial intelligence (AI) will be defined as "the subfield of Computer Science devoted to developing programs that enable computers to display behavior that can (broadly) be characterized as intelligent" (Russell & Norvig, 2010)[1]. There are usually considered to be two types of AI: "weak AI" (also referred as "narrow" or "applied" AI) and "strong AI". The former refers to machines that are neither self-aware nor intelligent in a generalizable way, but are able to complete well-defined yet complex (often human) tasks within narrow domains, like playing a game of Blackjack or recognizing and processing speech. The latter, strong AI, are machines that have reached a level of general intelligence that reflects, or possibly supersedes, that of human intelligence. This type of intelligence often includes the presence of conscious thought, although the nature of consciousness itself is still contested within cognitive science disciplines[2]. It should be noted that to date, scientists and academics are overall in consensus that we have not produced strong AI. Part of the reason for our failure to produce generalizable intelligence in machines is that over the past seventy-five years, the field of artificial intelligence has faced a number of problems due to its interdisciplinary nature. One of the disciplines that frequently poses a challenge for AI is philosophy, specifically the areas of philosophy of science and philosophy of mind.

Philosophy of science embodies a broad scope of knowledge related to our study of the world, and one of the ways it influences science is by examining and propagating scientific methodologies. In Western education, we are indoctrinated from an early age with a particular methodology of science: begin with an observation, acquire background knowledge, formulate a hypothesis, design a procedure, conduct an experiment, analyze its data, and conclude whether the original hypothesis has been falsified based on this data or whether it is supported by its results. This method of inquiry is favored by falsificationism, which is a theoretical approach to science that has helped shape modern scientific investigation. The scientific method we follow is so familiar to us that we rarely question it; however, it is

[1] Artificial intelligence can be approached from several different angles and disciplines, some of which take a human-centered approach and are concerned with modeling human intelligence, others of which take a rationalist approach and are concerned with modeling intelligent behavior that is not grounded in (and may exceed) human ability. For an overview of four approaches to AI, see Introduction in "Artificial Intelligence: A Modern Approach" (Russell, S. J. & Norvig, P., 2010). See also the "One Hundred Year Study on Artificial Intelligence" (A100) by Stone et al., 2016 for a discussion on how to define AI and a slightly different definition of intelligence.

[2] See Andy Clark's "Consciousness and the Meta-Hard Problem" in *Mindware* (Clark, 2014) for an overview of definitions of, approaches to, and challenges regarding the nature of consciousness. See also David Chalmers' *The Conscious Mind* (Chalmers, 1996) for another discussion on consciousness.

not the only—or even the first—proposed way to analyze and interpret the world according to philosophy, it is just the model that modern Western science favors.

The falsificationist model is a relatively recent concept in the general scheme of philosophy; philosopher Karl Popper first presented the model as a critique of the inductive scientific method in the twentieth century[3]. Under Popper's falsificationism, the falsifiability of a hypothesis is a necessary condition for its scientific merit. If a hypothesis cannot be disproven, or it is not understood what it would conceivably look like for it to be disproven, then there is no way to confirm or disconfirm with certainty whether it is valid science. Since the falsificationist model has been influential in the development of modern philosophy of science, it has become the primary way in which we demarcate the scientific from the non-scientific. One problem with this, however, is that some seemingly-scientific fields do not lend well to falsificationism, which threatens the very validity of the body of work. The field of artificial intelligence is one field that exemplifies this. Unfortunately, if the principles of falsificationism are what AI must adhere to for its developments to be considered scientific, we should be worried. As a consequence of our current, limited understanding of the mind—including the concepts of intelligence, thought, and personal identity—as well as our methods for delineating these concepts, it is unclear whether hypotheses concerning the design of artificial intelligence—particularly strong AI—are truly falsifiable. If this is the case, society's approach to and attitude towards future AI research and development regarding scientific methodology is likely in need of reevaluation.

When we bring into question the falsifiability of designing AI, we find there are those who claim our ability to design AI *is* falsifiable and, in fact, it has already been falsified. Historically, the best-known argument for the falsifiability of AI is contained in the 1972 published work, "Artificial Intelligence: A General Survey," commonly referred to as the Lighthill report. In his report, mathematician James Lighthill broke AI research down into three categories based on its goals:

1. Category A research aimed to automate specific human tasks by creating machines, known as Advanced Automatons, capable of performing the same tasks.
2. Category B research aimed to "bridge the gap" between Category A and Category C by unifying the field of AI through the physical process of Building Robots.
3. Category C research aimed to create Computer-based models of the Central Nervous System in order to investigate and understand human intellect from a neurobiological and psychological level.

[3] This argument was published in Popper's work "Conjectures and Refutations" (Popper, 1963) and has since been both embraced and criticized by philosophers and scientists. Whether or not Popper's model is correct or incorrect is beside the point at the moment. In order for the philosophical foundations of AI to be testable and falsifiable by the modern scientific method—which has undoubtedly been influenced by falsificationism—they must first be well-defined. This is true regardless of whether the AI is designed to model human level intelligence or beyond-human level intelligence.

Lighthill praised AI research in the areas of automation, such as information retrieval, speech recognition, and machine translation (Category A). He also praised research in the simulation of human physiology (Category C). However, he harshly critiqued research in language processing and robotics (Category B) (Lighthill, 1972). Essentially, he supported weak AI, but not strong AI. At the time it was published and in the two decades to follow, the Lighthill report was received by society as evidence that the field of AI was making little progress and had essentially failed to address its key issues such as the "no topic" problem[4], the frame problem[5] and combinatorial explosion[6]. Forty years ago, Lighthill's pessimistic outlook suggested the future of AI was bleak; some took it as an implication that the claim stating we could and would be able design strong AI was a false claim. That is, the proposition of generalizable artificial intelligence, as it was characterized by Lighthill, was not possible and therefore it had been falsified.

Since AI research in the areas of language processing and robotics (Category B) has progressed since the publication of the Lighthill report, it is clear that the future of AI is no longer regarded as a hopeless case. Much of the early work in artificial intelligence research was limited by physical symbol system and information processing approaches[7]. In the 1990s, the connectionist approach to AI revived previous interest in artificial neural networks and dynamic systems. Neural networks, by exhibiting graceful degradation, learning pattern recognition from experience, and implementing backpropagation algorithms, have since made progress in overcoming issues that plagued brittle early AI (Bostrom, 2014). In the past decade, the open-source movement and the field of data science have also accelerated the growth and capabilities of what we call 'AI' (Agrawal, McHale, & Oettl, 2017). In spite of these signs of progress, current attempts at AI still lack a developed description of the "intelligence" portion of "artificial intelligence": how cognition, thinking, and personal identity are described and evaluated in artificially intelligent systems, if they can be at all.

Because it is the case that intelligence and thought are concepts that are inherently tied to the investigation of artificial intelligence, it is fair to ask how we are to define them. Unfortunately, current work in AI produces no satisfying answers. One of the most well-known, though still contended,

---

[4] This is John Haugeland's term for when there is no relevant stereotype (something used to conceptualize knowledge in an organized way) to associate relevant information within a given situation. See Chapter 5: Real Machines in *Artificial Intelligence: The Very Idea* (Haugeland, 1985).

[5] Dennett coined this term in his paper "Cognitive Wheels: The Frame Problem of AI" (Dennett, 1985) to describe the challenge of logically representing the effects of actions without having to represent explicitly a large number of not-relevant effects that are intuitively obvious (to humans). The solution to the frame problem is knowing what knowledge is relevant at a given time.

[6] This is a term used to describe when there are more possibilities for a given thing than there is space or time to keep track of said thing and its possibilities, much less program them into AI. It is the reason why the frame problem is an issue for early AI.

[7] A more complete picture of the historical development of artificial intelligence can be found Introduction in "Artificial Intelligence: A Modern Approach" (Russell, S. J. & Norvig, P., 2010) and an overview of the shortcomings of the Physical Symbol System hypothesis is available in Chapter 2 of *Mindware* (Clark, 2014).

indications of intelligent thought within the field of AI is the Turing Test. Alan Turing, a mid-twentieth century computer scientist, developed the Turing Test as a method of validating machine intelligence. The test is Turing's own variation of the "imitation game;" it requires a human and a machine, both of which are hidden behind a curtain, and an investigator, who does not know the identities of either entity. If the investigator is unable to determine which entity is the machine and which is the human based upon their answers to questions, then the machine passes the Turing Test, at which point the entity's behavior sufficiently deems it capable of intelligent thought. If the investigator correctly detects which entity is the machine, then the machine does not pass the test, but that does not necessarily mean that it is incapable of intelligent thought (Turing, 1950).

The problems with the Turing Test and its position on intelligence and thought are multi-faceted. First, we must consider whether this *indication* of intelligent thought is an adequate *definition* of intelligence, assuming there is even a single correct definition of intelligence in the first place. Intelligence can, and has been, defined in a variety of ways, by multiple disciplines, and rarely are these definitions consistent[8]. Without a clear understanding of the conditions that constitute the notion of intelligence, it will be difficult—if not impossible—to falsify it. Of course, there are similar questions about what constitutes thought—is it purely computation or something more? —and how science must go about evaluating the existence or non-existence, as well as the depth, of thought in order to make it adequately falsifiable. The Turing Test is ultimately a behaviorist method which tries to simplify the problem of falsification by using observable behavior (the machine's responses) to make the determination of intelligence. However, behaviorism has problems of its own. In cognitive psychology, behaviorism was replaced by the information processing paradigm in the late twentieth century, and it is the information processing paradigm that has informed the majority of AI research. More recently, the paradigm has begun to shift towards the embodied perspective, more broadly known as situated cognition in philosophy[9]. Cognitive psychology recognizes that there is more to the mind than its behavior[10], so should this not also be the case for artificially-intelligent, thinking minds?

Another argument in support of falsifying the legitimacy of AI lies in weak AI that may be initially mistaken for strong AI. Examples of this are bots, such as customer service bots or chatbots, which humans interact with through web and mobile applications. Based on initial interactions with a bot,

---

[8] This definition often varies by discipline within the cognitive sciences. Cognitive psychology has taken multiple approaches to defining intelligence. Psychometric general (g) intelligence, Sternberg's triarchic theory, and Gardner's multiple intelligences theory are three ways in which intelligence has been defined. In philosophy, as Copeland points out in *What is Artificial Intelligence?* (Copeland, 2000) there is no cohesive definition of intelligence, artificial or otherwise, by which to evaluate AI.

[9] Embodiment maintains that the thinking mind is part of the body and, as such, is influenced by interactions with the external world. See Andy Clark's *Mindware* for further discussion of the embodied perspective.

[10] It should be noted that exactly what more constitutes the mind, besides behavior, and how to evaluate this is unclear and remains the subject of debate in cognitive psychology and other disciplines in cognitive science.

humans might assume that the customer service assistant on the IKEA website who asks "Can I help you?" through a chat window is actually another human at a service center somewhere in the world. However, after asking the virtual customer service assistant several conversational questions and receiving awkward, possibly nonsensical responses, humans will conclude that the assistant does not possess human-level intelligence, nor is it human, and therefore they have falsified the existence of strong AI in the bot[11]. There are multiple problems with this claim, all of which relate to the Turing Test. Like Turing's test for intelligent thought, falsifying the intelligence of a bot requires a behaviorist approach, which offers a worryingly incomplete picture of the mind, artificial or otherwise. Also, as stated above, passing the Turing Test is not a necessary condition for intelligence or the ability to think. That is: if the human correctly detects the bot, then the bot does not pass the test, but that does not necessarily mean that it is not intelligent or is not thinking. While it is true that we may all generally agree that the bot has not exhibited the behavior to have reached or exceeded human-level intelligence, we should be cautious about writing off entities as intelligence-lacking so quickly. Humans, for example, suffer from awkward and occasionally nonsensical responses, especially in instances where we lack sufficient knowledge about a topic. For example, consider monolingual foreign immigrants in a new country — their lack of behaviorally appropriate responses may be as strange as those of a machine, but it would be rather insulting to pronounce that the immigrants lack intelligence. Therefore, simple behaviorist approaches to falsifying AI are only somewhat helpful in distinguishing intelligence and certainly do not offer an adequate definition of the concept. For a more acceptable, valid scientific approach to falsifiability of AI, we should look beyond behaviorism and also consider other components of intelligence, such as consciousness.

Recently, academics in artificial intelligence and related disciplines have begun to voice some of these concerns. In 2015, AI academics and technology professionals signed an Open Letter on AI outlining the importance of investing in making AI more "robust" (e.g. capable) and beneficial to society. They followed this up with a longer report which maintains that AI has made significant progress and it will continue to make progress, so therefore it is valuable and even necessary to outline the areas of research that AI should prioritize in the coming years in order to solve current problems and preemptively address potential problems (Russell, Dewey, & Tegmark, 2015). Within this report, scholars raise questions of validity and verification in computer science research, specifically regarding strong AI. A major concern for AI is that it should be safely robust, meaning that it does only what we expect it to and does it well. In other words, it behaves as programmed. However, if we are to build AI that supposedly reaches (or exceeds) human level intelligence, it is going to gain knowledge and understanding in a way

[11] In fact, the bot is actually an example of weak AI, since it is programmed to do a defined, limited task, which is help people shop for furniture. Asking the IKEA customer service assistant what to have for breakfast is, unsurprisingly, not within the knowledge domain of its task.

that might not be known to us. Neural networks are—by the standards of strong AI—a primitive example of this today. With strong AI's ability to learn according to its own algorithms, we will not have a way other than its behavior to verify that it is learning what it is intended to learn, or determine really what kind of knowledge we are testing the AI for at all (Russell et al., 2015). The "brain" of the AI is a black box. Then, the question becomes: how do we know how and by what standard to evaluate what AI is doing and whether it is doing what it is supposed to when it is continually adapting and teaching itself? Is machine intelligence, no matter how we define it, falsifiable?

Finally, the semantic and conceptual concerns about intelligence and thought are connected to axiological concerns about both the scope and the falsifiability of personhood and personal identity. Assuming human level AI is achieved, under what circumstances, if any, should we grant it personhood? Currently, we do not even know how to validate or invalidate personhood in human beings. Common puzzles in philosophy of mind, cognitive science, and bioethics examine how we should classify amnesiacs and people in persistent vegetative state. The same issue of personhood's scope is just as relevant to artificial "persons". If AI were to become sentient enough to declare itself intelligent, thinking, or a person, we have no method by which to evaluate whether its claims are true or false. A related thought experiment in philosophy of mind that can be used to demonstrate this problem is the philosophical zombie. Proposed by David Chalmers, the philosophical zombie is an entity that is outwardly indistinguishable from a human being, but internally lacks the sentience and conscious experiences of humans. Chalmers uses the philosophical zombie to craft an argument against physicalism, since the conceivability of consciousness-lacking zombies implies that it is something immaterial, rather than material, which forms consciousness (Chalmers, 1996).[12] However, Chalmers' thought experiment can also be applied in this instance by replacing the philosophical zombie with an artificially intelligent agent: if the AI agent is a logically possible concept and if an AI agent were to claim that it is conscious and human, then this claim does not seem falsifiable. More generally, this is known as the "problem of other minds," and it is worth considering in the context of AI for its ethical implications, since a lack of falsifiability results in a lack of a clear way to determine whether potential strong AI is deserving of moral consideration.

So, what does this mean for the future of artificial intelligence in terms of falsificationist methodology? For the immediate future of AI, it is probably not a significant concern, since the possibility of strong AI is a long-term—but not unlikely—research goal, as scholars note in *Research Priorities for Robust and Beneficial Intelligence* (Russell et. al, 2015). The real concern is that our current methodology is not adequate for evaluating hypotheses regarding the scientific understanding and design

[12] Critics of Chalmers, such as Dennett, argue that zombies are logically and/or metaphysically impossible and therefore dismiss the p-zombie thought experiment. However, the physicalism versus dualism debate is not significant to the issue of falsifying zombie/AI's personhood that is presented in this paper.

of future artificial intelligence because falsifiability is not reachable. Either we should: (1) aim to better define the concepts in philosophy of mind that we attribute to artificial intelligence, like 'personal identity', 'thinking', and especially 'intelligence' itself, to understand how they can be falsified in order to make quicker and more meaningful progress in AI or (2) adapt our current scientific methodology to something more appropriate and inclusive than Popper's falsificationism, because perhaps there are concepts, such as aspects of AI, that simply cannot be falsified, yet are still scientifically valid and worthy of research[13]. In Popperian style, this much, at least, seems to get us a little closer to the truth about the process of falsifying artificial intelligence.

[13] Possible directions this could take include, but are not limited to (1) embracing Thomas Kuhn's pre-paradigm chaos as a way in which we can make valid scientific progress or (2) adopting Paul Feyerabend's rejection of epistemological methodology outlined in *Against Method* (Feyerabend, 1993) and instead favoring a more relativistic, anarchistic approach to doing science.

References

Agrawal, A., McHale, J., & Oettl, A. (2017). Finding needles in haystacks: artificial intelligence and recombinant growth (Rep.). http://www.nber.org/chapters/c14024.pdf

Artificial intelligence and life in 2030. (2016). One hundred year study on artificial intelligence. Report of the 2015 Study Panel. Stanford University. https://ai100.stanford.edu/2016-report

Bostrom, N. (2014). Superintelligence: paths, angers, strategies, 1st edition. Oxford University Press, Inc., New York, NY, USA.

Chalmers, D. (1996). The conscious mind.

Clark, A. (2014). Mindware: an introduction to the philosophy of cognitive science, 2nd edition. Oxford University Press.

Copeland, J. (2000). What is artificial intelligence. AlanTuring.net.

Dennett, D. (1984). Cognitive wheels: The frame problem of AI. In Christopher Hookway (ed.), Minds, Machines and Evolution. Cambridge University Press.

Feyerabend, P. (1993). Against method. Verso.

Haugeland, J. (1985). Artificial intelligence: The very idea. Cambridge, MA: MIT-Press.

Lighthill, J. (1972). Artificial intelligence: a general survey (Publication).

Popper, K. (1963). Science as falsification. In: Conjectures and Refutations, London: Routledge and Keagan Paul, 1963, pp. 33-39; from Theodore Schick, ed., Readings in the Philosophy of Science, Mountain View, CA: Mayfield Publishing Company, 2000, pp. 9-13.

Russell, S. and Norvig, P. (2010). Artificial intelligence: a modern approach, Englewood Cliffs, New Jersey: Prentice Hall, 3 ed., ISBN-10 0136042597.

Russell, S., Dewey, D., & Tegmark, M. (2015). Research priorities for robust and beneficial artificial intelligence. AI Magazine, 36(4), 105. doi:10.1609/aimag.v36i4.2577

Turing, A. M. (1950). Computing machinery and intelligence. Mind, 59, 433-460. http://dx.doi.org/10.1093/mind/LIX.236.433